Rate–Distortion Theory for FSQ

# 1. Problem Setup

In finite scalar quantization (FSQ), each dimension of a feature vector is quantized with a discrete set of levels. We want to minimize distortion subject to a bit-rate budget, ensuring the representation is efficient yet accurate.

# 2. High-Rate Quantization Theory

For scalar uniform quantization, distortion per axis with step size Δ is approximately D ≈ Δ²/12. If variance of dimension d is σ\_d² and we allocate b\_d bits, optimal allocation gives:  
D\_d ≈ c σ\_d² 2^(−2 b\_d).  
The total distortion is minimized by allocating more bits to higher-variance axes.

# 3. Bit Allocation Law

The classical water-filling solution gives:  
b\_d\* = (1/2) log₂( λ σ\_d² / c ),  
with λ chosen so that the sum of b\_d\* matches the target rate. This implies per-axis levels L\_d = 2^(b\_d\*), though FSQ allows non-powers of two.

# 4. Interpretation of [8, 6, 5, 5, 4]

For example, levels [8, 6, 5, 5, 4] correspond to bit allocations [3.0, 2.6, 2.3, 2.3, 2.0]. This distribution is justified if per-axis variances follow a decreasing pattern such as [2.3, 1.8, 1.2, 1.2, 0.9], naturally producing the observed allocation under the rate–distortion water-filling principle.

# 5. Practical Considerations

- Whitening or PCA ensures independence between axes so allocation theory applies cleanly.  
- Non-power-of-two levels are valid in FSQ; entropy coding compresses further.  
- Effective rate may be significantly lower than nominal rate if code usage is skewed.